Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Plant Biology

The emergence of whole genome association scans in barley

Robbie Waugh[1], Jean-Luc Jannink[2], Gary J Muehlbauer[3] and Luke Ramsay[1]

Barley geneticists are currently using association genetics to identify and fine map traits directly in elite plant breeding material. This has been made possible by the development of a highly parallel SNP assay platform that provides sufficient marker density for genome-wide scans and linkage disequilibrium-led gene identification. By leveraging the combined resources of the barley research and breeding sectors, marker-trait associations are being identified and a renewed interest has emerged in novel strategies for barley improvement. New database and visualization tools have been developed and statistical methods adapted from human genetics to account for complexities in the datasets. Exciting early results suggest that association genetics will assume a central role in establishing genotype-to-phenotype relationships.

Addresses
[1] Genetics, SCRI, Invergowrie, Dundee DD2 5DA, Scotland
[2] USDA-ARS, R.W. Holley Center for Agriculture and Health, Cornell University, Ithaca, NY 14853, USA
[3] Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108-6026, USA

Corresponding author: Waugh, Robbie (Robbie.Waugh@scri.ac.uk)

Current Opinion in Plant Biology 2009, **12**:218–222

This review comes from a themed issue on
Genome studies and molecular genetics
Edited by Masahiro Yano and Roberto Tuberosa

1369-5266/$ – see front matter
© 2009 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.pbi.2008.12.007

Introduction

Association genetics encompasses aspects of population and quantitative genetics that are needed to elucidate the history of a population in order to identify genotype to phenotype relationships on the basis of single and joint frequencies of observable genetic polymorphisms [1,2]. In the barley research and breeding communities there is increasing focus on association studies because they can directly analyze germplasm from contemporary breeding programs and consequently identify marker associations with breeding-relevant alleles. As the linkage disequilibrium (LD—see below) observed in such a study population arose outside and before the experiment through many rounds of meiosis and recombination, the resolution of observed associations is anticipated to be high. In this review, we describe what is known of the structure of LD in wild through cultivated barley, the design and implementation of genetic studies suitable to varying levels of LD, some of the bioinformatic and statistical tools and analyses being brought to bear, and finally some early successes and the directions they suggest for further study.

Linkage disequilibrium and association analysis

Alleles at two or more loci are said to be in LD if they are non-randomly co-inherited as determined by their individual and joint frequencies [3••]. Consequently, for two loci, the alleles at one locus are predictive of those present at the other. Given its dependence on allele frequencies, any measure of LD is population-specific. The forces that generate LD are also those that generate allele frequency changes, namely, mutation, drift and selection (Box 1) [3••,4]. The only process that systematically reduces LD is recombination, with LD between markers being expected to decay as a function of their recombination distance. This expectation is usually observed (Figure 1a), and forms the basis for decisions on the marker density required for effective association studies.

Short-range LD in wild species, landraces and cultivars

As a selfing species, LD in barley is predicted to be extensive [5,6]. Caldwell *et al.* [7•] examined short-range LD across a 212 kb sequence surrounding the barley *Hardness* (*Ha*) locus in cultivated, landrace and wild species (*H. vulgare* ssp. *spontaneum*) genepools. While highly significant association between paired sites extended across the region in the cultivated genepool, the landraces and wild species exhibited intermediate and rapid levels of decay, respectively. Rapid rates of intralocus LD decay were also observed in wild barley accessions by Morrell *et al.* [8] while elevated levels of intragenic LD between pairs of polymorphic sites in the *Bmy1* gene were observed in cultivated accessions [9]. Stracke *et al.* [10•] examined LD surrounding the *Hv-eIF4e* virus resistance locus by analyzing 83 SNPs over a 132 kb sequence in a diverse collection of cultivated and landrace accessions. In susceptible genotypes, LD broke down over distances of <1 cM whereas in resistant genotypes it extended completely across the region and by including data from genetically linked sites could be detected up to 5.5 cM away. Such observations are illustrative of specific blocks of extended LD (i.e. 'haplotype blocks') [11,12] being influenced by individual gene histories, including the effects of breeding and selection (as for *Hv-Eif4e*). Furthermore, the varying extent of LD observed in the different barley genepools

Box 1 Factors that generate LD

Mutation: Mutation generates LD because a mutant allele is associated initially only with those alleles present on the mutated chromosome.

Drift: Drift generates LD by sampling chromosomes at frequencies departing from their equilibrium expectations [3••].

Selection: For neutral loci, selection's effect depends on genetic hitch-hiking, whereby all alleles associated with a selected allele increase in frequency [3••]. This effect generates a signature of selection of extensive LD around the mutant selected allele but not the wild-type allele [4; and in barley, 10].

Population structure: Population structure is an important source of LD because each subpopulation contributes an excess of gametes carrying the alleles that it contains at high frequency.

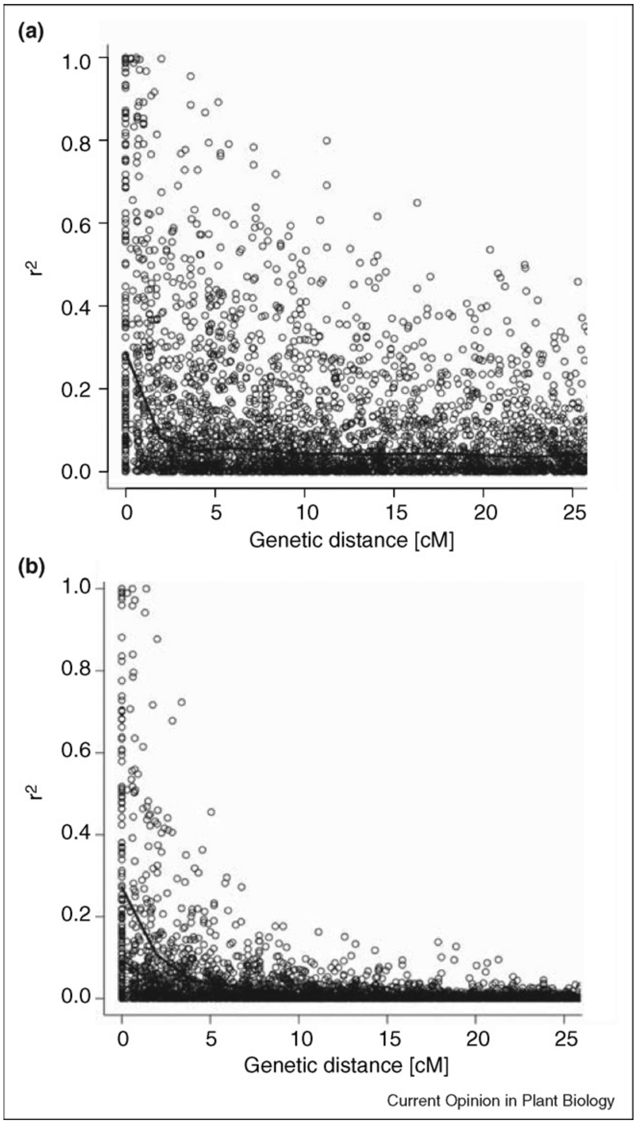
suggest that in gene discovery programs initial locus detection could be conducted in one genepool and fine-mapping and gene isolation in another [7•]. This strategy will require the same causal polymorphism to segregate and the relevant phenotype to be measurable in both elite and wild genepools.

Towards genome-wide association studies

It became clear from these short-range studies that the transition to genome-wide association mapping would require tools and genetic analyses pioneered in model species (e.g. human, maize, *Arabidopsis* [13]) to be adapted to inbreeding crop plants. International collaborators subsequently initiated the development of a highly multiplex unigene-based SNP assay platform for barley [14••]. They chose Illumina's oligo pool assay (OPA) [15] as a high-plex marker platform. Three pilot 1,536-plex OPAs have currently been evaluated and 3072 of the best performing assays compressed into two production barley OPAs (BOPA1 and 2) that are already available to the community. Approaching 3000 of these SNPs have been genetically mapped, enabling 'first generation' whole genome scans. An alternative marker platform, Diversity Array Technology (DArT), has also been developed and evaluated [16,17].

Using the first pilot OPA, Rostoks *et al.* [14••] provided initial evidence that the elite genepool could be effectively queried genome-wide. In a collection of 102 cultivars, LD extended from <1 to >10 cM, consistent with short-range LD observations and an earlier AFLP-based analysis of 146 modern spring barley cultivars [18•,19]. The elite population displays unique features resulting from a small number of founders [20] and its pseudo-outbreeding nature resulting from extensive intermating during breeding. On the basis of the observed LD and level of allelic variation in the elite genepool Rostoks *et al.* [14••] suggested that a few thousand bi-allelic SNPs may suffice for initial genome scans to discover marker-trait associations. This 'elite genepool strategy' has been adopted by two large association genetics programs in the US (BarleyCAP, http://barleycap.cfans.umn.edu) and UK (AGOUEB, http://www.agoueb.org). Both of these ambitious projects seek to increase marker utilization in barley breeding and genetics by mapping allelic variation in traits currently being manipulated in contemporary breeding programs. Other projects are embracing the multiple genepool concept, aiming to exploit the discriminatory LD observed in landrace and wild barley populations for fine mapping and gene identification (e.g. ExBarDiv: http://pgrc.ipk-gatersleben.de/barleynet/projects_exbardiv.php).

Figure 1



Decay of LD, as measured by r^2 against genetic distance in elite North American barley. In **(a)**, no adjustment for population structure has been made. In **(b)**, the r^2 value shown is a partial r^2 adjusted for structure by multiple correlation including 20 principle component eigenvectors. Lines are LOWESS regressions of r^2 on genetic distance.

Statistical analysis issues

Data from non-experimental populations create both analysis challenges and opportunities. Firstly, population structure (differential relatedness among individuals) can generate spurious association results. Analyses that account for populations composed of subpopulations, for lines related through pedigree, and for both exist [21–23]. Explicit modelling of subpopulations can be performed [24], but model-free methods (e.g. principal components analysis) are also effective [25,26]. Adjustments for structure generally lead to lower but more appropriate LD estimates (Figure 1b). Secondly, higher marker densities frequently result in higher numbers of missing and erroneous data points. Methods developed in human genetics allow missing marker data to be imputed from surrounding loci [27]. Barley data have provided the means to test imputation methods in a crop, creating options to reduce genotyping demands and error rates [28]. Thirdly, at current marker densities, it is feasible to identify blocks of SNPs that are strongly associated and these so-called 'haplotype blocks' may be useful units of analysis to predict the phenotype (e.g. [29,30]).

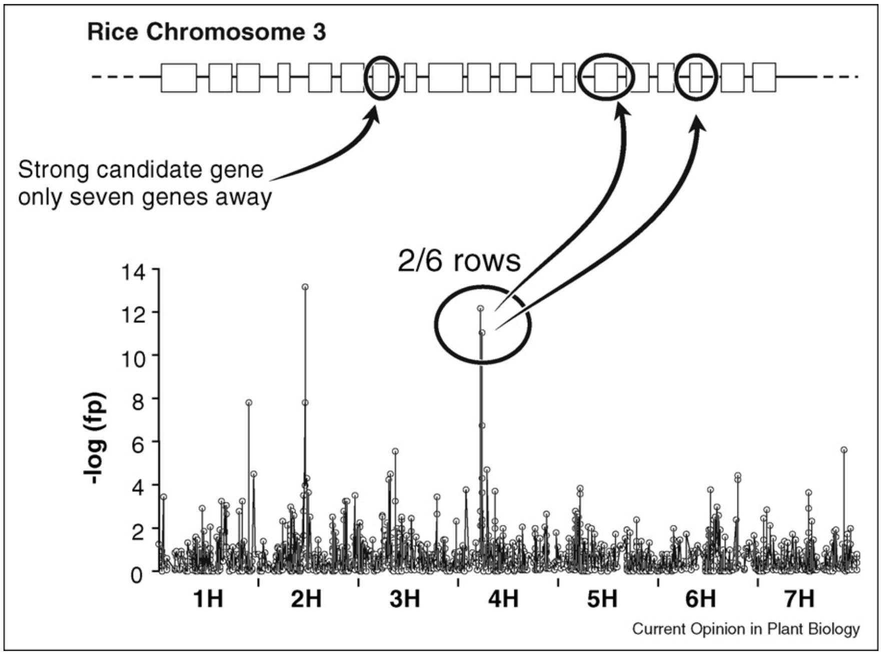
Data storage and visualization

The use of high-plex marker assay platforms on thousands of accessions quickly generates large amounts of information that require new bioinformatics and data handling tools. The Germinate data-base schema [31] was designed to handle all data types related to plant genetic resources and has been extended to genotypic data. BarleyCAP and AGOUEB data are being housed in Germinate-derived databases along with genotypic, phenotypic and pedigree information. 'The Hordeum Toolbox' (http://www.hordeumtoolbox.org/) developed in Barley CAP also includes phenotype and genotype datasets that can be searched and sorted according to user preference. Novel visualization tools such as Flapjack (http://www.scri.ac.uk/research/genetics/platformtechnologies/bioinformatics/software) facilitate interpretation of genotypic datasets allowing researchers and breeders to interrogate aligned genotypes in a user-friendly environment. Such tools are a vital component of effective and practical translational genomics programs that aspire towards predictive crop improvement.

Some early successes

The initial survey of Rostoks *et al.* [14••] gives some indication of the potential of association genetics in elite barleys. As a surrogate for Mendelian traits the authors assigned putative map locations for 43 out of 85 unmapped SNP loci through association with those already mapped. They also found a strong association in a region on barley chromosome 5HL with winter/spring growth habit, despite correspondence of the trait with population structure. Encouragingly the region showing association coincided with a cluster of *Arabidopsis CBF*

Figure 2



Outline strategy for candidate *int-c* gene identification using rice (or *Brachypodium*) as a genomic model. A whole genome association scan identified strongly associated gene-based SNP markers (bottom panel: *x* and *y* axes give respectively marker genetic positions and negative log of the no-association null hypothesis p-value). These were blasted against the rice genome revealing adjacent orthologous sequences on a segment exhibiting extensive conserved synteny between barley and rice. Surveying annotations of the surrounding genes identified a strong candidate gene that was confirmed by resequencing an allelic series of *int-c* alleles (Ramsay, in preparation).

gene homologues that are key regulators of the cold acclimation signalling pathway [32]. This potential to map Mendelian traits by association has been confirmed recently by fine mapping a morphological character, rachilla hair length, to 5HL using a panel of 192 individuals and 4600 SNPs (Comadran J, unpublished).

The relevant scale for mapping resolution is the number of candidate genes that could plausibly be associated with a phenotype, which depends on LD decay but also on gene density and the relationship between genetic and physical distance [33,34]. In certain regions, the discrimination achieved through association mapping can delineate sufficiently small genomic regions to allow strong candidate genes to be assigned through conserved synteny with genomic models. In our laboratory we have used this feature to identify SNPs in genomic regions on 2HL and on 4HS associated with ear-type (two or six row ear), a trait known to be largely controlled by two genes, *vrs1* and *int-c*. Through synteny with rice, we found that the most significantly associated SNP on 2HL was seven genes away from the recently cloned *vrs1* [35•]. Similarly for *int-c* on 4HS the two most highly associated SNPs were two genes apart with a very strong candidate gene a further seven genes away. Re-sequencing this candidate in stocks of induced mutants at *int-c* confirmed this candidate as the causal gene (Figure 2; Ramsay L, unpublished). Thus, association mapping using a panel of 192 lines has enabled the isolation of *int-c*. By comparison, a segregating progeny of 13 093 gametes was necessary to clone *vrs1* [35•].

Prospects

Results to date have shown that for many traits there will be considerable merit in investing further in genome-wide association mapping as both a locus and gene discovery tool. However, further research remains necessary for optimizing both population sizes (and substructure) and density of molecular markers in order to afford routine success. The assembly and release of community-wide mapping panels, comprised initially of elite genetic materials accompanied by genotypic information, is already under discussion and will address these issues. Community panels will allow individual researchers to focus on collecting specific phenotypic information, which can be queried remotely to reveal marker-trait associations, and a possible route towards gene isolation. While the multiple genepool concept is potentially very powerful and robust data are required to confirm it as a viable strategy, preliminary indications are promising. For example, highly significant associations between DArT markers and rust resistance in wild material has been reported [36], although no genes have yet been identified. Studies in these genepools are important because they offer the very real possibility of identifying novel allelic variation that may be of considerable value to future crop improvement. Finally, it has become increasingly apparent that the depth and quality of the OPA SNP data is shining a very powerful light on the relations between individual cultivars. As a result, end-user tailored software tools are urgently required to facilitate innovative exploitation of the data and to drive new genetic strategies focused on practical crop improvement.

Acknowledgements

RW and LR would like to acknowledge SA LINK 302/BB/D522003/1 Association Genetics of Elite UK Barleys for funding to support the development of LD-mapping in the elite UK barley genepool. GJM and J-LJ acknowledge funding from USDA-CSREES-NRI for supporting the parallel Barley CAP focusing on the US elite genepool. Special thanks go to Prof T Close and colleagues for their tireless efforts in developing the pilot and production barley OPAs.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Rafalski JA: **Novel genetic mapping tools in plants: SNPs and LD-based approaches**. *Plant Sci* 2002, **162**:329-333.

2. Botstein D, Risch N: **Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease**. *Nat Genet* 2003, **33**:228-237.

3. •• Slatkin M: **Linkage disequilibrium – understanding the evolutionary past and mapping the medical future**. *Nat Rev Genet* 2008, **9**:477-485.
Excellent review of the causes and applications of LD in single or multiple populations.

4. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R *et al.*: **Genome-wide detection and characterization of positive selection in human populations**. *Nature* 2007, **449**:913-918.

5. Nordborg M: **Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization**. *Genetics* 2000, **154**:923-929.

6. Flint-Garcia SA, Thornsberry JM, Buckler ES: **Structure of linkage disequilibrium in plants**. *Annu Rev Plant Biol* 2003, **54**:357-374.

7. • Caldwell KS, Russell J, Langridge P, Powell W: **Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species *Hordeum vulgare***. *Genetics* 2006, **172**:557-567.
A detailed study of short-range LD surrounding the hardness locus in barley. The much greater extent of LD in cultivated than wild barley suggested a two-tiered approach to fine mapping through association using cultivated germplasm to identify candidate regions and wild germplasm to further localize causal polymorphisms. Analysis also suggested that selection and genome structure affected LD patterns.

8. Morrell PL, Toleno DM, Lundy KE, Clegg MT: **Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization**. *Proc. Natl. Acad. Sci U S A* 2005, **102**:2442-2447.

9. Malysheva-Otto LV, Roeder MS: **Haplotype diversity in the endosperm specific β-amylase gene *Bmy1* of cultivated barley (*Hordeum vulgare* L.)**. *Mol Breeding* 2006, **18**:143-156.

10. • Stracke S, Presterl T, Stein N, Perovic D, Ordon F, Graner A: **Effects of introgression and recombination on haplotype structure and linkage disequilibrium surrounding a locus encoding Bymovirus resistance in barley**. *Genetics* 2007, **175**:805-817.
Fascinating work analyzing effects of introgression and selection on haplotype extent and LD decay at a locus affecting resistance to barley yellow mosaic virus. Extended haplotype conservation and LD were observed in germplasm carrying resistance alleles compared to germplasm carrying susceptibility alleles. Selection was inferred to have increased resistance allele frequency faster than recombination could decay LD.

11. Jeffreys AJ, Kauppi L, Neumann R: **Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex**. *Nat Genet* 2001, **29**:217-222.

12. Haubold B, Kroymann J, Ratzka A, Mitchell-Olds T, Wiehe T: **Recombination and gene conversion in a 170-kb genomic region of *Arabidopsis thaliana***. *Genetics* 2002, **161**:1269-1278.

13. Zhao K, Aranzana MaJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P *et al.*: **An Arabidopsis example of association mapping in structured samples**. *PLoS Genetics* 2007, **3**:e4.

14. •• Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF *et al.*: **Recent history of artificial outcrossing facilitates whole genome association mapping in elite inbred crop varieties**. *Proc Natl Acad Sci U S A* 2006, **103**:18656-18661.

This article reports foundational research towards whole-genome association scans in barley. On the basis of 1391 SNP genotyped across 102 elite mostly European barley lines, the authors estimate rates of LD decay and show that Mendelian traits can be effectively mapped. They estimate that several hundred to a few thousand SNP will enable effective association mapping and provide useful targets for marker assisted selection.

15. Fan J-B, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P *et al.*: **Highly parallel SNP genotyping**. *Cold Spring Harb Symp Quant Biol* 2003, **68**:69-78.

16. Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A: **Diversity arrays technology (DArT) for whole-genome profiling of barley**. *Proc Natl Acad Sci U S A* 2004, **101**:9915-9920.

17. Wenzl P, Li H, Carling J, Zhou M, Raman H, Paul E, Hearnden P, Maier C, Xia L, Caig V *et al.*: **A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and agricultural traits**. *BMC Genomics* 2006, **7**:206.

18. • Kraakman ATW, Niks RE, van den Berg PMMM, Stam P, Van Eeuwijk FA: **Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars**. *Genetics* 2004, **168**:435-446.

Reports on the earliest example of a whole genome association scan using phenotypic data derived from breeding and variety evaluation activity rather than basic research activity, showing the potential to leverage applied crop improvement activities for genomic studies.

19. Kraakman ATW, Martnez F, Mussiraliev B, van Eeuwijk FA, Niks RE: **Linkage disequilibrium mapping of morphological, resistance, and other agronomically relevant traits in modern spring barley cultivars**. *Mol Breeding* 2006, **17**:41-58.

20. Fischbeck G: **Diversification through breeding**. In *Diversity in Barley.* Edited by Bothmer R, van Hintum T, Knupffer H, Sato K. Amsterdam: Elsevier; 2003:29-52.

21. Kennedy BW, Quinton M, van Arendonk JAM: **Estimation of effects of single genes on quantitative traits**. *J Anim Sci* 1992, **70**:2000-2012.

22. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: **Association mapping in structured populations**. *Am J Hum Genet* 2000, **67**:170-181.

23. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB *et al.*: **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness**. *Nat Genet* 2006, **38**:203-208.

24. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data**. *Genetics* 2000, **155**:945-959.

25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies**. *Nat Genet* 2006, **38**:904-909.

26. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis**. *PLoS Genetics* 2006, **2**:e190.

27. Scheet P, Stephens M: **A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase**. *Am J Hum Genet* 2006, **78**:629-644.

28. Marchini J, Howie B, Myers S, McVean G, Donnelly P: **A new multipoint method for genome-wide association studies by imputation of genotypes**. *Nat Genet* 2007, **39**:906-913.

29. Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP: **Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes**. *Am J Hum Genet* 2004, **75**:35-43.

30. Zhao HH, Fernando RL, Dekkers JCM: **Power and precision of alternate methods for linkage disequilibrium mapping of quantitative trait loci**. *Genetics* 2007, **175**:1975-1986.

31. Lee JM, Davenport GF, Marshall D, Ellis THN, Ambrose MJ, Dicks J, van Hintum TJL, Flavell AJ: **GERMINATE. A generic database for integrating genotypic and phenotypic information for plant genetic resource collections**. *Plant Physiol* 2006, **139**:619-631.

32. Szucs P, Skinner JS, Karsai I, Cuesta-Marcos A, Haggard KG, Corey AE, Chen THH, Hayes PM: **Validation of the *VRN-H2/VRN-H1* epistatic model in barley reveals intro length variation in *VRN-H1* may account for a continuum of vernalization sensitivity**. *Mol Genet Genomics* 2007, **277**:249-261.

33. Kunzel G, Korzun L, Meister A: **Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints**. *Genetics* 2000, **154**:397-412.

34. Kunzel G, Waugh R: **Integration of microsatellite markers into the translocation-based physical RFLP map of barley chromosome 3H**. *Theor Appl Genet* 2002, **105**:660-665.

35. • Pourkheirandish M, Wicker T, Stein N, Fujimura T, Komatsuda T: **Analysis of the barley chromosome 2 region containing the six-rowed spike gene *vrs1* reveals a breakdown of rice-barley micro collinearity by a transposition**. *Theor Appl Genet* 2007, **114**:1357-1365.

Paradigmatic use of micro-synteny between rice chromosome 4 and barley chromosome 2H to clone barley *vrs*1. Nevertheless, the rice orthologue *Vrs*1 was found on chromosome 7 indicating a breakdown in microsynteny and suggesting that *vrs*1 transposed in barley after its separation from rice.

36. Steffenson BJ, Olivera P, Roy JK, Jin Y, Smith KP, Muehlbauer GJ: **A walk on the wild side: mining wild wheat and barley collections for rust resistance genes**. *Aust J Agric Res* 2007, **58**:532-544.